Pediatric Surgery International (2022) 38:1057–1066
https://doi.org/10.1007/s00383-022-05133-y

ORIGINAL ARTICLE

Paediatric surgical trials, their fragility index, and why to avoid using it to evaluate results

Arne Schröder[1] · **Oliver J. Muensterer**[2,3] · **Christina Oetzmann von Sochaczewski**[3,4]

Accepted: 24 April 2022 / Published online: 7 May 2022

Abstract

Background The fragility index has been gaining ground in the evaluation of comparative clinical studies. Many scientists evaluated trials in their fields and deemed them to be fragile, although there is no consensus on the definition of fragility. We aimed to calculate the fragility index and its permutations for paediatric surgical trials.

Methods We searched pubmed for prospectively conducted paediatric surgical trials with intervention and control group without limitations and calculated their (reverse) fragility indices and respective quotients along with posthoc-power. Relationships between variables were evaluated using Spearman's ρ. We also calculated S values by negative log transformation base-2 of P values.

Results Of 516 retrieved records, we included 87. The median fragility index was 1.5 (interquartile range: 0–4) and the median reverse fragility index was 3 (interquartile range: 2–4), although they were statistically not different (Mood's test: $\chi^2 = 0.557$, df = 1, $P = 0.4556$). P values and fragility indices were strongly inversely correlated ($\rho = -0.71$, 95% confidence interval: -0.53 to -0.85, $P < 0.0001$), while reverse fragility indices were moderately correlated to P values ($\rho = 0.5$, 95% confidence interval: 0.37–0.62, $P < 0.0001$). A fragility index of 1 resulted from P values between 0.039 and 0.003, which resulted in S values between 4 and 8.

Conclusions Fragility indices, reverse fragility indices, and their respective fragility quotients of paediatric surgical trials are low. The fragility index can be viewed as no more than a transformed P value with even more substantial limitations. Its inherent penalisation of small studies irrespective of their clinical relevance is particularly harmful for paediatric surgery. Consequently, the fragility index should be avoided.

Keywords Reverse fragility index · Fragility quotient · Paediatric surgery · Uninformative metric · S value

Introduction

The fragility index dates back to 2014 [1] and has been widely popularised in many specialties, [2–5] including paediatric surgery [6]. The fragility index relies on the concept of statistical significance, calculated by Fisher's exact test, and its overturn by adding events to the group with the smallest number until statistical significance collapses. The required number of additional events is then defined as the fragility index.

Several extensions to the fragility index have been described: (1) The fragility quotient, division of the fragility index by sample size, to provide a relative measure reflecting trial sample size [7]. (2) The reverse fragility index, which is calculated by the same iterative process as in the fragility index, but aims to add events until the statistical significance collapses. Thereby, it reports the number of events that

✉ Christina Oetzmann von Sochaczewski
c.oetzmann@gmail.com

[1] Klinik für Kinder- und Jugendmedizin, Klinikum Dortmund, Dortmund, Germany

[2] Kinderchirurgische Klinik und Poliklinik im Dr. von Haunerschen Kinderspital, Ludwig-Maximilians-Universität München, München, Germany

[3] Klinik und Poliklinik für Kinderchirurgie, Universitätsmedizin der Johannes-Gutenberg-Universität Mainz, Mainz, Germany

[4] Sektion Kinderchirurgie der Klinik und Poliklinik für Allgemein, Viszeral, Thorax- und Gefäßchirurgie, Universitätsklinikum Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

would have transformed statistically insignificant results into statistically significant ones [8]. (3) The fragility index for meta-analyses[9] and network meta-analyses [10] that allow a similar assessment, not only of trials, but their synthesis in meta-analyses.

Statistical details matter, as the controversy [11] around the readdressing [12] of the concept of posthoc-power has shown. Post-hoc power is just a transformed *P* value, [13] but this flawed [14, 15] concept has been put forward as a remedy for underpowered studies in surgery [16]. Therefore, the concept of fragility has been widely embraced due to its appealing simplicity [17]. It is marketed to simplify a complex issue, analysing trial relevance and robustness, by transformation into a single metric [18].

However, several drawbacks come with the fragility index: it can only be used to evaluate dichotomous outcomes and thereby relies on Fisher's exact test irrespective of the test used to evaluate the results in the included trials [19]. Another issue is interpretation of the fragility index: trials can either be evaluated as fragile or as robust, but there is no clear definition of both terms [18, 20]. Consequently, terming a result fragile or robust includes a "catchy connotation" [18] that might be used to include "spin"—altered presentation of the facts—in the presentation of results of a trial, which might then disconnect its reporting from the actual results [19].

We, therefore, explored the fragility indices of paediatric surgical trials to assess the implications of the results for our specialty.

Methods

We searched PubMed for paediatric surgical trials published through the 31st December 2019 without limitations for time, language or document type. The search was conducted on the 22nd of January 2020 and produced 516 eligible records. Following title and abstract screening, 139 records remained that were subjected to full-text evaluation. From these, 87 publications including 243 eligible comparisons were included in our analysis. Screening was conducted by two researchers in parallel and disagreements were solved by discussion and consensus. Eligible for inclusion were only prospectively conducted trials with an intervention and a separate control group. Included publications had at least one comparison with a dichotomous outcome. Included comparisons were separated by primary outcome, secondary outcome, and the remaining comparisons including those in which primary and secondary outcomes were not stated. Based on the extensions of the fragility index beyond a 1:1 allocation ratio, [5, 10, 21] we did not limit our analysis to trials with such an allocation ratio. Only two trials that did not fulfil this requirement were included. For transparency, the list of the included studies with outcomes and extracted numbers can be freely accessed from a repository [22].

Statistical analysis was conducted using R [23] (version 3.5.3) with its generic stats4-package if not stated otherwise. Fragility and reverse fragility indices were calculated as described in the introduction using the fragility index-package (version 0.1.0) [24]. The stepwise calculation of the (reverse) fragility index has been described and visualised in detail elsewhere [25]. Fragility and reverse fragility quotients were calculated by dividing them by the trial's sample size, [7] but multiplied with 100 to avoid excessively small numbers [26]. Data are presented as medians with interquartile ranges. Correlation analyses were conducted using Spearman's ρ, whose 95% confidence intervals were calculated using bootstrapping with 10,000 repetitions [27, 28] via the spearman.ci-function from the RVAideMemoire-package (version 0.9-75) [29]. Posthoc-power was calculated using the power2 × 2-function from the exact2 × 2-package (version 1.6.5) [30]. Medians were compared using Mood's test from the RVAideMemoire-package (version 0.9–75) [31]. *P* values were Shannon-transformed by calculating their negative base-2 logarithm [32].

Results

We included 87 different publications in our analysis, the majority of which (48/87) were published between 2013 and 2019 (Fig. 1A). The Journal of Pediatric Surgery was the most frequent publication venue with 31% (27/87), followed by Annals of Surgery with 9% (8/87), and Pediatric Surgery International with 7% (6/87) of all 32 included journals (Fig. 1B). The most frequent subject areas were gastrointestinal surgery in 23% (20/87), followed by paediatric urology in 16% (14/87), and paediatric appendicitis in 15% (13/87) (Fig. 1B).

The median fragility index of included comparisons was 1.5 (interquartile range: 0–4) (Fig. 2A) and had a median fragility quotient of 1.89% (interquartile range: 0–4.87%) (Fig. 2B). For the reverse fragility index, the median was 3 (interquartile range: 2–4) (Fig. 2C) with a corresponding median fragility quotient of 4.03% (interquartile range: 2.25–6.67%) (Fig. 2D). Dropping the two non-1:1-allocation ratio studies did neither change the median reverse fragility quotient nor its interquartile range. However, it resulted in a subtly altered interquartile range for the median fragility quotient of the reverse fragility index with an interquartile range from 2.31 to 6.67%.

Fragility indices and *P* values were highly inversely correlated ($\rho = -0.71$, 95% confidence interval: −0.53 to −0.85, $P < 0.0001$) (Fig. 3A). In addition, the fragility indices were fairly correlated to patient numbers in the trials ($\rho = 0.25$, 95% confidence interval: 0.03–0.47, $P = 0.0323$) (Fig. 3B).

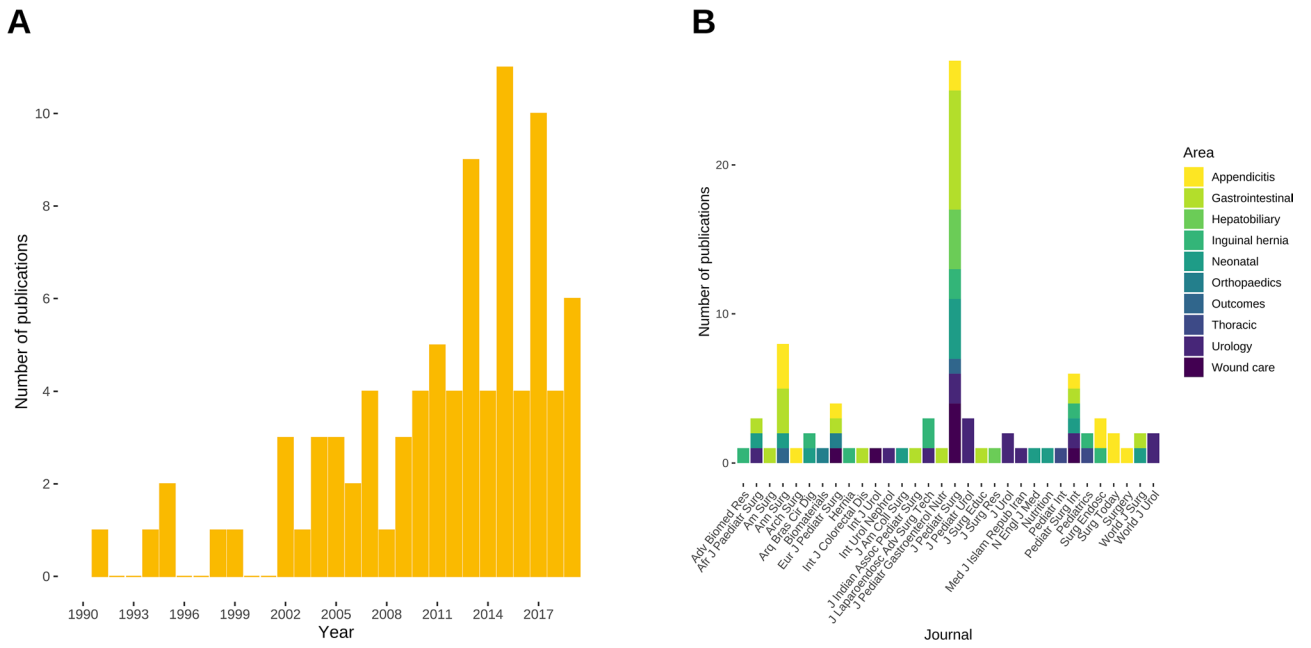

Fig. 1 Details of the included studies. Publication year of the included studies (**A**) and publication venue by subject area (**B**)

Likewise, reverse fragility indices were moderately correlated to *P* values ($\rho = 0.5$, 95% confidence interval: 0.37–0.62, $P < 0.0001$) (Fig. 3C) and weakly correlated to patient numbers in the trials ($\rho = 0.17$, 95% confidence interval: 0.007–0.32, $P = 0.0303$) (Fig. 3D).

There was no difference in medians between the fragility and the reverse fragility index ($\chi^2 = 0.557$, df = 1, $P = 0.4556$), whereas the median reverse fragility quotient was larger than the median fragility quotient ($\chi^2 = 11.035$, df = 1, $P = 0.0009$).

P values and posthoc-power were highly inversely correlated ($\rho = -0.88$, 95% confidence interval: −0.84 to −0.92, $P < 0.0001$) (Fig. 4A). Consequently, fragility index and posthoc-power were highly correlated ($\rho = 0.83$, 95% confidence interval: 0.68–0.93, $P < 0.0001$), too (Fig. 4B). Likewise, posthoc-power and the reverse fragility index were moderately inversely correlated ($\rho = -0.66$, 95% confidence interval: −0.53 to −0.77, $P < 0.0001$) (Fig. 4C).

The same fragility index of 1 was calculated from trials with a sample size between 30 and 243 patients and from *P* values ranging from 0.003 to 0.039. Corresponding *S* values of these *P* values are between 4.68 and 8.38 (Table 1), representing a large difference in probabilities that is not covered by the fragility index.

Discussion

Following its "invention", the fragility index has been widely embraced by several disciplines that examined the robustness of trials in their specialty and unanimously concluded that the vast majority of trials in their field were fragile [17]. Earlier propositions to evaluate trial fragility were accompanied by a theoretical framework [33, 34]. On the contrary, the first description of the fragility index [1] was more a description than an thorough evaluation without adequate statistical simulation supporting its arguments, despite a necessity to do so [35].

Our results of a median fragility index of 1.5 with an interquartile range from 0 to 4 are similar to preceding analyses in children: in paediatric critical care, the median fragility index was 2 with an interquartile range of 1–6, [36] it was 0 with an interquartile range from 0 to 2 for preoperative androgen stimulation for hypospadias surgery, [37] and 3 with an interquartile range from 0.75 to 4.25 in paediatric appendicitis [6]. Recent results in adults are not different either: For irritable bowel syndrome, the median fragility index was 6 with an interquartile range from 0 to 25.25, [38] for proximal humerus fractures, the median fragility index was 1 with an interquartile range from 0 to 3 [39]. These results are overall similar to what has been reported before in spine surgery, otolaryngology, ophthalmology, sports surgery, and orthopaedic surgery, which all have a similar median fragility index of 2 or 3 [20].

These similar results could be expected due to the similar process in study design. Based on clinical judgement or preceding exploratory research, an effect size is defined based on the clinically relevant minimal difference [40, 41]. This is the starting point for a sample size calculation that should be able to demonstrate this difference: usually with a statistical power (1-β) of 80% and the typical

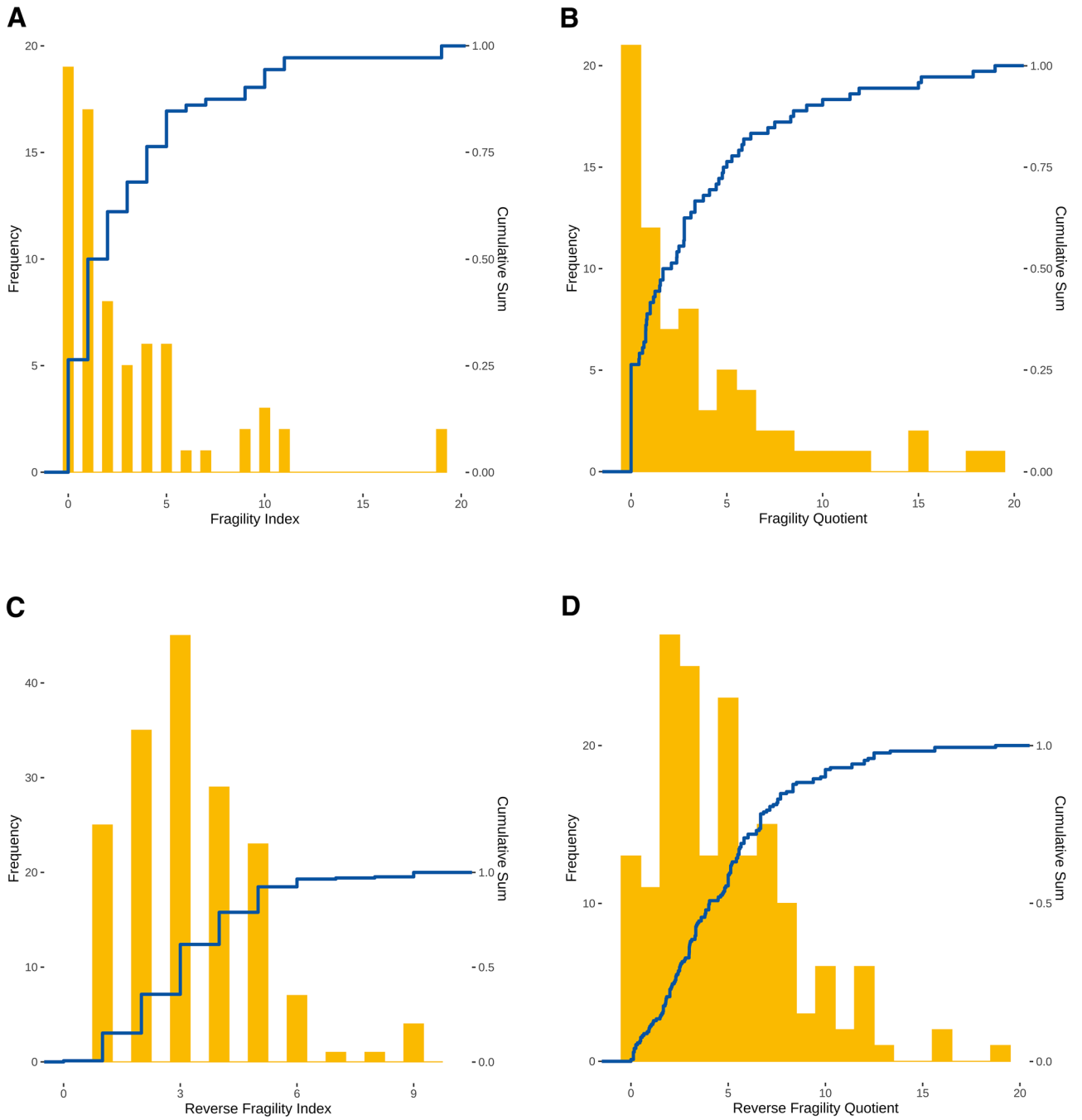

Fig. 2 Distribution of fragility indices and fragility quotients. Histograms and cumulative sums of fragility indices (**A**), fragility quotients (**B**), reverse fragility indices (**C**), and reverse fragility quotients (**D**). Fragility quotients are displayed as percent to avoid excessively small numbers

α-level of 5% [41]. If the sample size is determined this way, the resulting *P* value will be slightly below the conventional cutoff of 0.05. In this context, it is important to remember the definition of the *P* value: "The probability that the chosen test statistic would have been at least as large as its observed value if every model assumptions were correct, including the test hypothesis" [42].

If the sample size of the study in question was designed to demonstrate a minimally clinically important difference with as few patients as possible, the resulting *P* value will usually not be very small: the input data are not planned to be extremely different from the null-hypothesis, but only different enough to demonstrate the effect based on conventional significance definitions. Thus, all trials will be

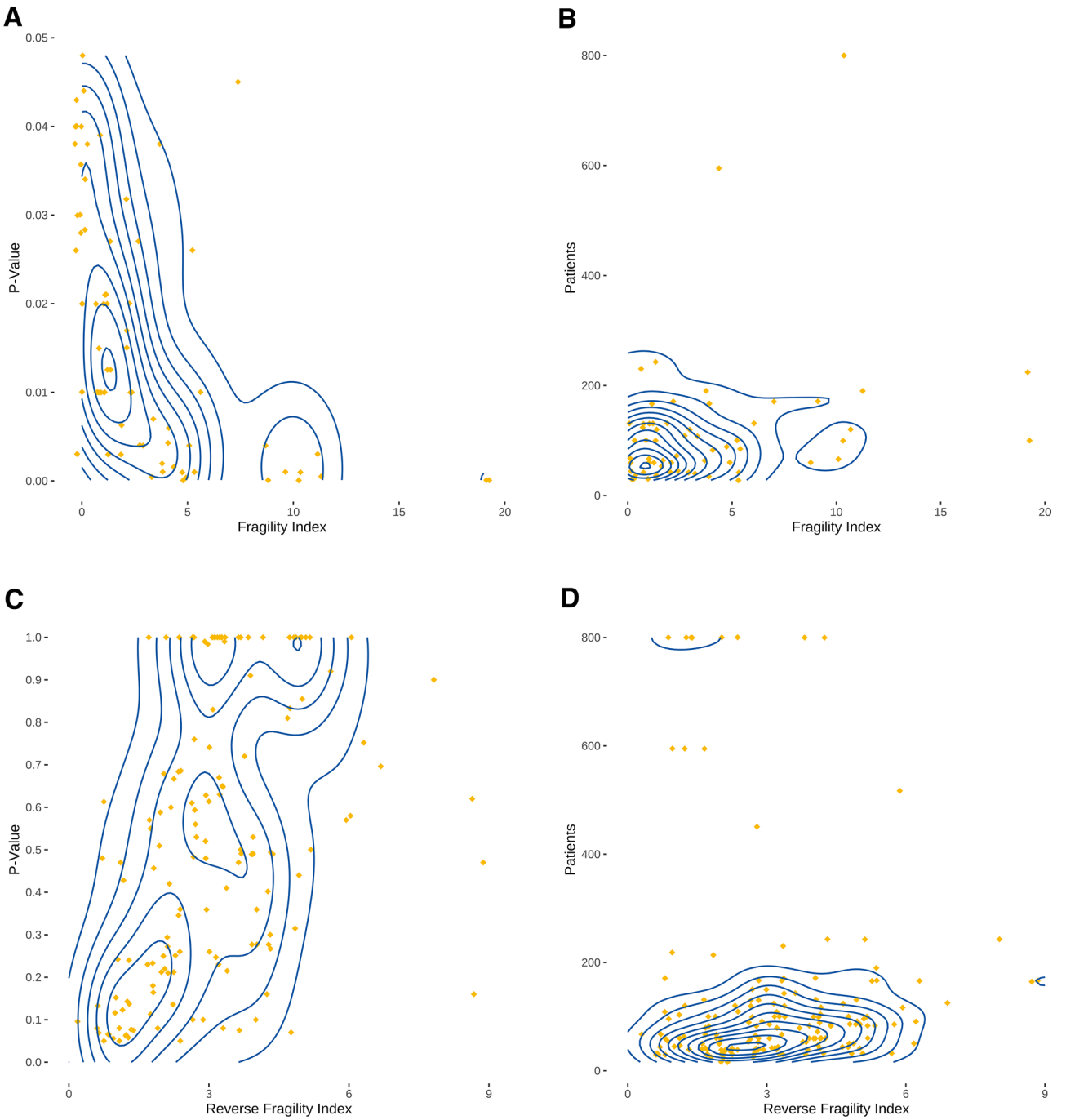

Fig. 3 Correlation between fragility indices, *P* values, and sample size. Diamonds represent individual comparisons. Correlation analyses were conducted using Spearman's ρ. **A** Fragility indices and *P* values were inversely highly correlated ($\rho = -0.71$, 95% confidence interval: −0.53 to −0.85, $P < 0.0001$). **B** Fragility indices and patient numbers were fairly correlated ($\rho = 0.25$, 95% confidence interval: 0.03–0.47, $P = 0.0323$) **C** Reverse fragility indices and *P* values were moderately correlated ($\rho = 0.5$, 95% confidence interval: 0.37–0.62, $P < 0.0001$). **D** Reverse fragility indices and patient numbers were weakly correlated ($\rho = 0.17$, 95% confidence interval: 0.007–0.32, $P = 0.0303$)

[43] and are fragile, [20] unless the trialist has unlimited funds to expand the sample size towards infinity [44]. The fragility index ignores these principles of trial design [18, 45].

The *P* value is inherently linked to sample size: if the effect size is not exactly zero, there will always be a statistically significant *P* value, given a sufficiently large sample [40, 46]. Fisher's exact test calculates its *P* value by the sum

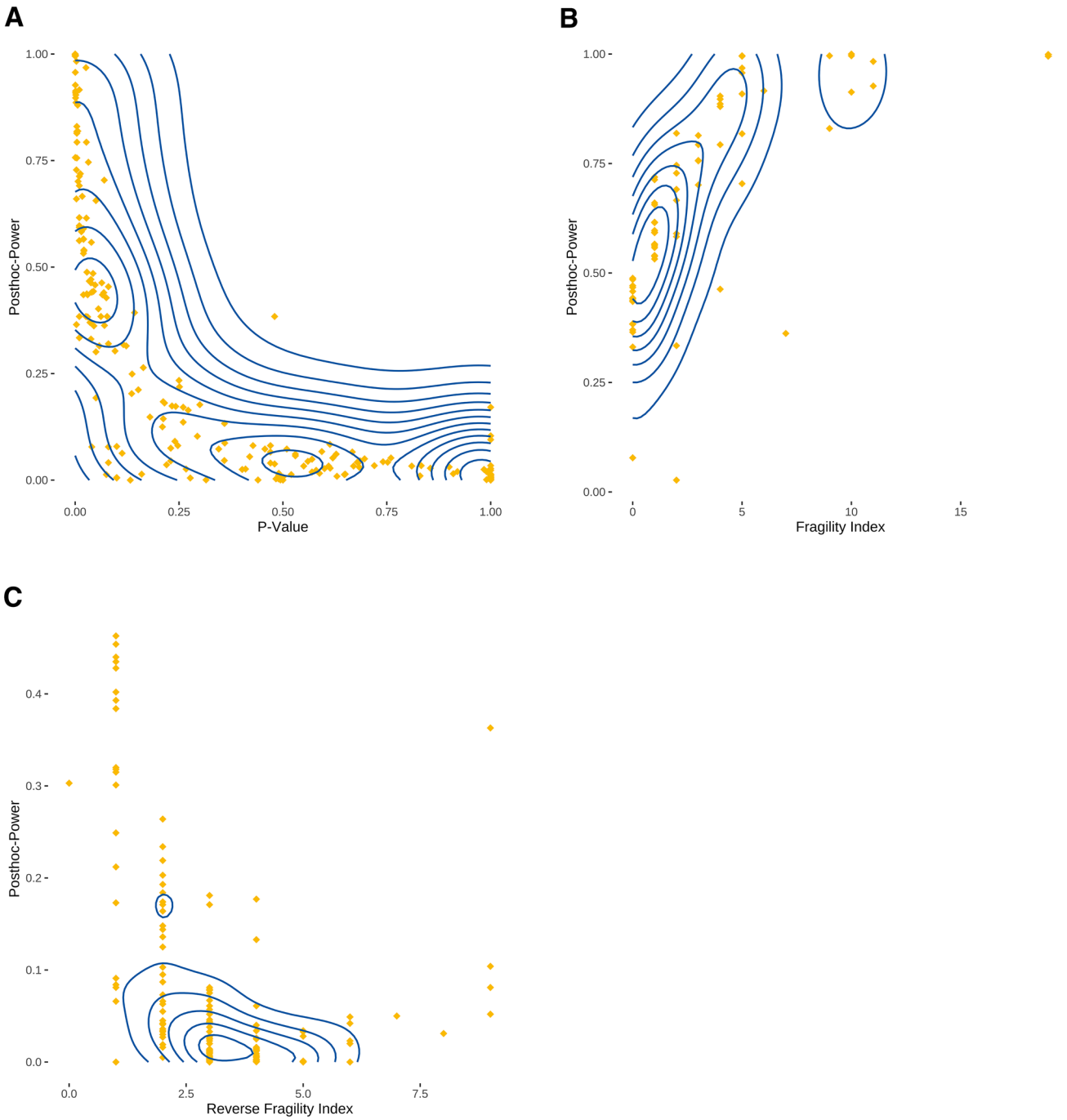

Fig. 4 Correlation between posthoc-power, *P* values, and fragility indices. Diamonds represent individual comparisons. Correlation analyses were conducted using Spearman's ρ. **A** *P* values and posthoc-power were inversely strongly correlated ($\rho = -0.88$, 95% confidence interval: −0.84 to −0.92, $P < 0.0001$). **B** Posthoc-power and fragility indices were highly correlated ($\rho = 0.83$, 95% confidence interval: 0.68–0.93, $P < 0.0001$). **C** Posthoc-power and reverse fragility indices were moderately inversely correlated ($\rho = -0.66$, 95% confidence interval: −0.53 to −0.77, $P < 0.0001$)

of two by two tables that are equal or more extreme than the observed ones [40, 47]. Consequently, an increasing sample size in a study will also increase the number of tables that are more extreme than observed, thereby reducing the *P* value. This relationship explains the correlation between fragility, reverse fragility indices, and patient numbers in the trials. This relationship could be expected based on statistical simulation [40, 48, 49]. This aspect is of particular relevance for paediatric surgery: in addition to the challenges of all surgical specialties when conducting trials, [50]

Table 1 Same fragility index can be derived from many *P* values

Study population	Fragility index	*P* value	*S* value
120	1	0.039	4.68
130	1	0.027	5.21
60	1	0.021	5.57
60	1	0.021	5.57
230	1	0.02	5.64
167	1	0.02	5.64
36	1	0.02	5.64
124	1	0.015	6.06
100	1	0.0125	6.32
100	1	0.0125	6.32
243	1	0.01	6.64
131	1	0.01	6.64
131	1	0.01	6.64
67	1	0.01	6.64
36	1	0.01	6.64
30	1	0.01	6.64
42	1	0.003	8.38

Shannon-transformation (negative log-transformation with base 2) of *P* values for all fragility indices of 1 from the included trials. The bits of information, *s*, may be interpreted as the probability of getting only heads in s unbiased coin tosses

paediatric surgery is further limited by the low incidence of congenital anomalies [51]. Small studies are penalised by the fragility index as it inevitably takes less events with a changed outcome to overturn statistical significance [45, 49].

This issue has been explained in detail based on an example of two hypothetical trials, [52] with the same *P* value of 0.02, but with 1 of 100 and 200 and of 4000 patients in the treatment group compared to 9 of 100 and 250 of 4000 patients in the control group who experienced a negative event. The relative risk of 0.11 in the treatment group is much more relevant than the relative risk of 0.8 in the larger trial, but the fragility index favours the larger trial [52]. The graphical depiction using consonance curves [32] emphasises this: the relative risk of the smaller trial is farther from the null and thus represents a stronger effect compared to a larger trial with the same *P* value [53].

It has been specifically discussed for paediatric surgery using the example of the highly effective foetal endoscopic tracheal occlusion for isolated congenital diaphragmatic hernia, which reports a massive clinical difference of a tenfold risk of death in the control group with conventional treatment [53]. Nevertheless, this study could be determined fragile based on its fragility index of 3, which is caused by the small sample size that penalises the large clinical effect.

This effect is rooted in the derivation of the fragility index from the *P* value. Therefore, it has been named simply a "repackaged" *P* value [40] or a "surrogate parameter" for the *P* value [48]. The close relationship between the *P* value and the fragility index has already been described in early reports, [1–3] but it took some years until their close connection due to the derivation of the fragility index from the *P* value was demonstrated using statistical simulation [40, 48, 49]. Consequently, we observed exactly the pattern that would be expected from simulation studies: a strong inverse correlation between *P* values and fragility indices. This is also in line from what has been reported in orthopaedic trauma surgery, [54] irritable bowel syndrome, [38] paediatric appendicitis, [6] paediatrics, [55] and many more [17].

Reverse fragility indices did not behave differently: only the direction of the effect changed, because reverse fragility does not aim to remove statistical significance, but to reach it [8, 24]. Thus, we observed a strong positive correlation between *P* values and reverse fragility indices in our data set, similar to preceding assessments in other clinical specialties [8, 54]. The close relationship to the *P* value may further be depicted using posthoc-power, which simply is a function of the *P* value [13, 56]. The assessment of *P* values and posthoc-power demonstrated exactly the relationship between them predicted by statistical simulation before [11, 15]. Consequently, posthoc-power, a function of the *P* value, and both fragility indices and reverse fragility indices were correlated, similar to the results in orthopaedic surgery [48].

Due to the derivation of the fragility index from the *P* value, Porco & Lietman concluded that "Fragility retains all the problems of the *P* value, with none of the usefulness—and frankly, none of the charm" [44]. Much has been written about the problems of the *P* value, [42, 57] but the usefulness of the original version of the *P* value compared to the derived (reverse) fragility index may easily be demonstrated using Shannon-transformation. The negative base-2 logarithm of the *P* value yields the *S* value, which can ease interpretation of statistical results by nonprobabilistic measures of information [32].

The *S* value provides bits of information by which *P* values can be described by comprehensible information with known probabilities, such as a coin toss with a non-manipulated coin: [32] An *S* value of 5 would hence represent the same evidence against the null hypothesis as would five coin tosses with a non-manipulated coin that showed all heads. The same fragility index may be calculated from a widely different range of *P* values in our data and in simulation studies [48, 49]. The *S* value then exposes the weakness of this concept: the fragility index of 1 in our data corresponds to a wide range of differences in probabilities of which the highest has 91% more evidence against the null hypothesis than the smallest. The fragility index in contrast would just lump them together and label them as fragile. Consequently, the fragility index obscures important distinctions: the differences in probabilities are obvious at first sight if a sequence of four heads in four coin tosses would be equal

to eight coin tosses with eight heads in a row, because they result in the same fragility index of one.

A limitation of our analysis is the inherent restriction to trials with dichotomous outcomes that is rooted within the calculation of the fragility index using Fisher's exact test [40, 48, 49, 52]. This precludes the analysis of all numeric or time to event outcomes. Although the latter might be transformed to dichotomous data if it is simplified to the simple question if the endpoint has been reached or not, but this would be inadequate. It would strip the analysis of the important aspect of time that has passed: a patient would most likely agree that there is a relevant difference if survival is 6 months compared to 5 years. So do we and, therefore, did not conduct such analyses. Nevertheless, our search strategy seemed exhaustive enough based on similar results of the fragility index compared to other fields in both adults [1, 2, 5, 20, 38, 39] and children [6, 36, 37, 55].

Just recently, Caldwell et al. proposed an extension of the fragility index to continuous outcomes by an iterative approach: they conducted a Welch-test and changed the data point with the mean closest to the mean of the control group to the mean of the control group, which is repeated until the *P* value of the Welch-test becomes nonsignificant [58]. Apart from the fact that this method has only been tested using simulated data sets, it still retains the problem that the metric is based on the *P* value and thus inherits the problems of the conventional fragility index.

Taking into account the critique of the fragility index that it may rely on inappropriately rare outcome modifications, [59] an extension of the traditional fragility index has been proposed [60]. This extension can be generalised beyond the 2×2 table and also precludes unlikely modifications by taking into account the distribution of the outcome resulting in "sufficiently likely" outcome modifications [60]. However, due to its mathematical complexity, this method is beyond the scope of this manuscript. The same group has suggested that the fragility index may be used for sample size calculations in clinical trials [61]. They used two examples of coronary artery disease to illustrate their suggestions: with an estimated fragility index of 15, the sample size of one trial increased by 45% and with a fragility index of 25, the sample size of the other trial increased by 89% [61].

Apart from the financial aspect that the trialist would require much more money to conduct such a trial, it also raises an ethical issue: a trial is designed to establish its aim with as few patients as possible to minimise potential harm to those randomised to the inferior intervention [62]. It would thus be unethical to further randomise patients to an intervention that is known—due to the mandatory intermediate review—to be inferior, just to achieve certain levels of the fragility index [62]. For this specific aim, the adaptive trial design has been developed to avoid both underpowered trials and randomising patients to futile treatments [63]. Such a trial would always be considered fragile based on the fragility index alone, [62] which emphasises that it may not be used to determine trial sample size.

Besides this quite technical line of argument, the use of the fragility index in paediatric surgery is rather limited per se. It could have only been used on results of (randomised) controlled, prospectively conducted trials with a dichotomous outcome and thus not applicable to the vast majority of research data in paediatric surgery. Moreover, our study addresses only one of the many potential errors that can be made during the scientific process: [64] The wrong interpretation of the *P* value and its derivatives are potentially harmful, but the focus on them should not obscure other problems, such as poor study design and data collection or the "hunt for significance" [65]. They also would inevitably lead to poor results or as it had been pointed out in the context of meta-analyses already in 2001 with the famous proverb "garbage in, garbage out" [66]. Nonetheless, the still ongoing discussion about the fragility index might emphasise the point of critical assessment of research methods per se: "Perhaps the most valuable contribution of the fragility index is that its shortcomings and inconsistencies have inspired scientists to reflect deeply on what it truly means […]," [52] or more pointed "So it is perhaps odd that, while we allow a doctor to conduct surgery only after years of training, we give SPSS® (SPSS, Chicago, IL) to almost anyone"[67].

Fragility indices, reverse fragility indices, and their respective fragility quotients of paediatric surgical trials are low. The fragility index can be viewed as no more than a transformed *P* value with even more substantial limitations. Its inherent penalisation of small studies irrespective of their clinical relevance is particularly harmful for paediatric surgery. Consequently, the fragility index should be avoided.

Author contributions Study conception and design: AS, CO. Data acquisition: AS, CO. Analysis and data interpretation: AS, CO, OM. Drafting of the manuscript: AS, CO. Critical revision: OM. All authors have read and approved the manuscript. All authors agree with the submission.

Funding Open Access funding enabled and organized by Projekt DEAL. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We received no financial support for the research, authorship or publication of this article.

Availability of data and materials The data sets generated and analysed during the current study are available in the Zenodo repository (https://doi.org/10.5281/zenodo.4883231) [22].

Declarations

Competing interests The authors declare no competing interests.

Conflict of interest We have nothing to declare.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C et al (2014) The statistical significance of randomized controlled trial results is frequently fragile: a case for a fragility index. J Clin Epidemiol 67:622–628
2. Ridgeon EE, Young PJ, Bellomo R, Mucchetti M, Lembo R, Landoni G (2016) The fragility index in multicenter randomized controlled critical care trials. Crit Care Med 44:1278–1284
3. Mazzinari G, Ball L, Serpa Neto A, Errando CL, Dondorp AM, Bos LD et al (2018) The fragility of statistically significant findings in randomised controlled anaesthesiology trials: systematic review of the medical literature. Brit J Anaesth 120:935–941
4. Tignanelli CJ, Napolitano LM (2019) The fragility index in randomized clinical trials as a means of optimizing patient care. JAMA Surg 154:74–79
5. Bertaggia L, Baiardo Redaelli M, Lembo R, Sartini C, Cuffaro R, Corrao F et al (2019) The fragility index in peri-operative randomised trials that reported significant mortality effects in adults. Anaesthesia 74:1057–1060
6. Robinson T, Al-Shahwani N, Easterbrook B, VanHouwelingen L (2020) The fragility of statistically significant findings from randomized controlled trials in pediatric appendicitis: a systematic review. J Pediatr Surg 55:800–804
7. Ahmed W, Fowler RA, McCredie VA (2016) Does sample size matter when interpreting the fragility index? Crit Care Med 44:e1142–e1143
8. Khan MS, Fonarow GC, Friede T, Lateef N, Khan SU, Anker SD et al (2020) Application of the reverse fragility index to statistically nonsignificant randomized clinical trial results. JAMA Netw Open 3:e2012469
9. Atal I, Porcher R, Boutron I, Ravaud P (2019) The statistical significance of meta-analyses is frequently fragile: definition of a fragility index for meta-analyses. J Clin Epidemiol 111:32–40
10. Xing A, Chu H, Lin L (2020) Fragility index of network meta-analysis with application to smoking cessation data. J Clin Epidemiol 127:29–39
11. Schröder A, Oetzmann von Sochaczewski C (2020) On the Difference between a-priori and observed statistical power—a comment on "statistical power and sample size calculations: a primer for pediatric surgeons." J Pediatr Surg 55:203–205
12. Bababekov YJ, Stapleton SM, Mueller JL, Fong ZV, Chang DC (2018) A proposal to mitigate the consequences of type 2 error in surgical science. Ann Surg 267:621–622
13. Hoenig JM, Heisey DM (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 55:19–24
14. Althouse AD, Chow ZR (2019) Comment on "post-hoc power: if you must, at least try to understand." Ann Surg 270:e78–e79
15. Althouse AD (2021) Post hoc power: not empowering just misleading. J Surg Res 259:A3-6
16. Bababekov YJ, Hung Y-C, Hsu Y-T, Udelsman BV, Mueller JL, Lin H-Y et al (2019) Is the power threshold of 0.8 applicable to surgical science?—empowering the underpowered study. J Surg Res 241:235–239
17. Holek M, Bdair F, Khan M, Walsh M, Devereaux PJ, Walter SD et al (2020) Fragility of clinical trials across research fields: a synthesis of methodological reviews. Contemp Clin Trials 97:106151
18. Chaitoff A, Zheutlin A, Niforatos JD (2020) The fragility index and trial significance. JAMA Intern Med 180:1554
19. Lobo DN (2019) Fragility, Spin, and interpretation of randomized clinical trials. Crit Care Med 47:486–488
20. Dettori JR, Norvell DC (2020) How fragile are the results of a trial? the fragility index. Global Spine J 10:940–942
21. Grammatikopoulou MG, Nigdelis MP, Theodoridis X, Gkiouras K, Tranidou A, Papamitsou T et al (2021) How fragile are Mediterranean diet interventions? A research-on-research study of randomised controlled trials. BMJ Nutr Prev Health 4:115–131
22. Schröder A, Muensterer OJ, Oetzmann von Sochaczewski C (2022) Paediatric surgical trials, their fragility index, and why to avoid using it to evaluate trials. Dataset for the evaluation of fragility indices in paediatric surgical trials. https://doi.org/10.5281/zenodo.4883231
23. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2019.
24. Johnson KW, Rappaport E, Shameer K, Glicksberg BS, Dudley JT. fragilityindex: an R package for statistical fragility estimates in biomedicine. Preprint. Bioinformatics; 2019.
25. Lin L (2021) Factors that impact fragility index and their visualizations. J Eval Clin Pract 27:356–364
26. Schröder A, Muensterer OJ, Oetzmann von Sochaczewski C (2021) Meta-analyses in paediatric surgery are often fragile: implications and consequences. Pediatr Surg Int 37:363–367
27. Baumgart J, Deigendesch N, Lindner A, Muensterer OJ, Schröder A, Heimann A et al (2020) Using multidimensional scaling in model choice for congenital oesophageal atresia: similarity analysis of human autopsy organ weights with those from a comparative assessment of aachen minipig and pietrain piglets. Lab Anim 54:576–587
28. Oetzmann von Sochaczewski C, Tagkalos E, Lindner A, Lang H, Heimann A, Muensterer OJ (2019) Technical aspects in esophageal lengthening: an investigation of traction procedures and suturing techniques in swine. Eur J Pediatr Surg 29:481–484
29. Hervé M. RVAideMemoire: Testing and Plotting Procedures for Biostatistics. 2020.
30. Fay MP, Hunsberger SA, Nason M, Gabriel E, Lumbard K. Exact2x2: exact tests and confidence intervals for 2 × 2 tables. 2020.
31. Llano López LH, Melonari P, Gehring S, Schreiner D, Grucci S, Pérez Araujo S et al (2021) Point-of-care multiplex-PCR enables germ identification in haemolytic uremic syndrome 94 h earlier than stool culture. Eur J Clin Microbiol Infect Dis 40:643–645
32. Rafi Z, Greenland S (2020) Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Med Res Methodol 20:244

33. Feinstein AR (1990) The unit fragility index: an additional appraisal of "statistical significance" for a contrast of two proportions. J Clin Epidemiol 43:201–209
34. Walter SD (1991) Statistical significance and fragility criteria for assessing a difference of two proportions. J Clin Epidemiol 44:1373–1378
35. Morris TP, White IR, Crowther MJ (2019) Using simulation studies to evaluate statistical methods. Stat Med 38:2074–2102
36. Matics TJ, Khan N, Jani P, Kane JM (2019) The fragility of statistically significant findings in pediatric critical care randomized controlled trials*. Pediatr Crit Care Med 20:258–262
37. Li B, Kong I, McGrath M, Farrokhyar F, Braga LH (2021) Evaluating the literature on preoperative androgen stimulation for hypospadias repair using the fragility index—can we trust observational studies? J Pediatr Urol 17:661–669
38. Williams MO, Sedarous M, Dennis B, Dlamini V, Nwaiwu O, Nguyen L et al (2021) The fragility of randomized placebo-controlled trials for irritable bowel syndrome. Neurogastroenterol Motil 33:14166
39. Kyriakides PW, Schultz BJ, Egol K, Leucht P (2021) The fragility and reverse fragility indices of proximal humerus fracture randomized controlled trials: a systematic review. Eur J Trauma Emerg Surg. https://doi.org/10.1007/s00068-021-01684-2
40. Carter RE, McKie PM, Storlie CB (2017) The fragility index: a P-value in sheep's clothing? Eur Heart J 38:346–348
41. Staffa SJ, Zurakowski D (2020) Statistical power and sample size calculations: a primer for pediatric surgeons. J Pediatr Surg 55:1173–1179
42. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN et al (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31:337–350
43. Niforatos JD, Zheutlin AR, Chaitoff A, Pescatore RM (2020) The fragility index of practice changing clinical trials is low and highly correlated with P-values. J Clin Epidemiol 119:140–142
44. Porco TC, Lietman TM (2018) A fragility index: handle with care. Ophthalmology 125:649
45. Acuna SA, Sue-Chue-Lam C, Dossa F (2019) The fragility index— P values reimagined flaws and all. JAMA Surg 154:674
46. Sullivan GM, Feinn R (2012) Using effect size—or why the *P* value is not enough. J Grad Med Educ 4:279–282
47. Fay MP (2010) Confidence intervals that match Fisher's exact or Blaker's exact tests. Biostatistics 11:373–374
48. Reito A, Raittio L, Helminen O (2019) Fragility index, power, strength and robustness of findings in sports medicine and arthroscopic surgery: a secondary analysis of data from a study on use of the Fragility Index in sports surgery. PeerJ 7:e6813
49. Condon TM, Sexton RW, Wells AJ, To M-S (2020) The weakness of fragility index exposed in an analysis of the traumatic brain injury management guidelines: a meta-epidemiological and simulation study. PLoS ONE 15:e0237879
50. Cook JA (2009) The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. Trials 10:9
51. Gelijns AC, Ascheim DD, Parides MK, Kent KC, Moskowitz AJ (2009) Randomized trials in surgery. Surgery 145:581–587
52. Potter GE (2020) Dismantling the fragility index: a demonstration of statistical reasoning. Stat Med 39:3720–3731
53. Schröder A, Muensterer OJ, von Oetzmann Sochaczewski C (2021) The fragility index may not be ideal for paediatric surgical conditions: the example of foetal endoscopic tracheal occlusion. Pediatr Surg Int 37:967–969
54. Forrester LA, McCormick KL, Bonsignore-Opp L, Tedesco LJ, Baranek ES, Jang ES et al (2021) Statistical fragility of surgical clinical trials in orthopaedic trauma. JAAOS Glob Res Rev 5:e20.00197
55. Matics T, Khan N, Jani P, Kane J (2017) The fragility index in a cohort of pediatric randomized controlled trials. JCM 6:79
56. O'Keefe DJ (2007) Brief report: post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: sorting out appropriate uses of statistical power analyses. Commun Methods Meas 1:291–299
57. Goodman S (2008) A dirty dozen: twelve P-value misconceptions. Semin Hematol 45:135–140
58. Caldwell J-ME, Youssefzadeh K, Limpisvasti O (2021) A method for calculating the fragility index of continuous outcomes. J Clin Epidemiol 136:20–25
59. Walter SD, Thabane L, Briel M (2020) The fragility of trial results involves more than statistical significance alone. J Clin Epidemiol 124:34–41
60. Baer BR, Gaudino M, Charlson M, Fremes SE, Wells MT (2021) Fragility indices for only sufficiently likely modifications. Proc Natl Acad Sci USA 118:e2105254118
61. Baer BR, Gaudino M, Fremes SE, Charlson M, Wells MT (2021) The fragility index can be used for sample size calculations in clinical trials. J Clin Epidemiol 139:199–209
62. Stensland KD, Daignault-Newton S, Skolarus TA (2021) Designing lean, efficient clinical trials is an ethical imperative: the fragility index should not be used in the design of randomized clinical trials. Urolo Oncol 39:738–739
63. Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV et al (2018) Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med 16:29
64. Brown AW, Kaiser KA, Allison DB (2018) Issues with data and analyses: errors, underlying themes, and potential solutions. Proc Natl Acad Sci USA 115:2563–2570
65. Amrhein V, KornerNievergelt F, Roth T (2017) The earth is flat significance thresholds and the crisis of unreplicable research. PeerJ 5:3544
66. Egger M, Smith GD, Sterne JAC (2001) Uses and abuses of meta-analysis. Clin Med 1:478–484
67. Vickers A (2005) Interpreting data from randomized trials: the scandinavian prostatectomy study illustrates two common errors. Nat Rev Urol 2:404–405